情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-BIO-21 No.16
2010/6/19

Computational identification of discriminating features of pathogenic and symbiotic type III secreted effector proteins

Koji Yahara[†] Ying Jiang[†] Takashi Yanagawa[††]

Type III secretion systems (T3SS) deliver bacterial proteins, or "effectors", into eukaryotic host cells, inducing physiological responses in the hosts. Effector proteins have been considered virulence factors of pathogenic bacteria, but T3SSs have now been found in symbiotic bacteria as well. Whether any physicochemical difference exists between the two types of effectors remains unknown. In this work, we combined computational statistical and machine-learning methods to find the physicochemical differences. The most discriminating set of features in a dataset of physicochemical features was determined using generalized Bayesian information criteria and kernel logistic regression. Classification performance was examined using a support vector machine. Interdependence among the most discriminating features was explored by graphical modeling, and the most discriminating region was investigated by sliding window analysis.

病原細菌と共生細菌の III 型分泌装置の エフェクタータンパク質を区別する特徴は何か？

矢原耕史[†] 姜英[†] 柳川堯[††]

近年、細菌がホスト細胞に送り込むエフェクタータンパク質が、病原細菌と共生細菌の双方に存在することが注目されているが、その差異は未だ明らかでない。本研究では、両者の差異を司る物理化学的特徴セットをカーネルロジスティック回帰の情報量基準によって抽出し、その判別性能を SVM によって評価し、さらにその相互依存関係及び最も特徴的な領域を、グラフィカルモデリングと sliding window 解析によって明らかにした。

1. Introduction

Type III secretion systems (T3SS) are complex secretion machines that deliver bacterial proteins called effectors into eukaryotic host cells through an injectisome during infection (1, 2). T3SS-secreted effector proteins induce physiological responses in their hosts, such as cytoskeletal rearrangement to promote bacterial attachment and invasion, interference with cellular trafficking processes, cytotoxicity (2), induction of apoptosis of macrophages (3), disruption of tight junctions (4), and microtubule destabilization (5). These effector protein functions are considered causes of virulence in pathogenic bacteria such as *Yersinia* species (spp.), *Chlamydia* spp., *Salmonella* spp., *Shigella* spp., and enteropathogenic *Escherchia coli*. However, T3SSs are also found in symbiotic bacteria (6, 7), and a genome analysis of a Chlamydia-related symbiont of free-living amoebae suggests that the origins of T3SSs may be unrelated to virulence (8).

Common features of T3SS effector proteins in pathogenic and symbiotic bacteria can be identified by computational methods (9, 10). While T3SS effector proteins were originally not thought to share any common features (11), recent studies using machine-learning approaches have identified commonalities in the N-terminus of effectors, mainly in amino acid composition. One study (9) analyzed both pathogenic and symbiotic T3SS effector proteins, and found a signature in the N-terminus that is taxonomically universal and conserved.

The symbiotic T3SS effector proteins, however, have different functions than the pathogenic effectors. Symbiotic effectors of rhizobia, for example, modulate host-plant reactions, that lead to the formation of functional nodules (12, 13). Putative effector proteins of the tsetse fly endosymbiont, *Sodalis glossinidius*, specifically facilitate the host cell cytoskeletal rearrangements necessary for bacterial entry, although the number of genes encoding effector proteins is smaller in the symbiotic regions than in the homologous islands in pathogenic bacteria (14). Homologs of the symbiotic regions are also found in endosymbionts of grain weevils, *Sitophilus oryzae* and *S. zeamais,* in which T3SS genes are suggested to function during a specific stage of weevil development (14). Even if the signature amino acid sequence in the N-terminus is conserved among pathogenic and symbiotic T3SS effector proteins, these functional differences exist. We were interested in

[†] Division of Biostatistics, Kurume University School of Medicine
久留米大学医学研究科バイオ統計学群
[††] Biostatistics Center, Kurume University
久留米大学バイオ統計センター

finding the physicochemical differences between pathogenic and symbiotic T3SS effector proteins that might be responsible for these functional differences.

In this work, we combined computational statistical and machine-learning approaches to address this issue. From a dataset of physicochemical features prepared from pathogenic and symbiotic T3SS effector proteins, the most discriminating set of features was determined using generalized Baysian information criteria and kernel logistic regression. Classification performance using the identified discriminating features was examined using support vector machine (SVM). The results clearly showed differences in amino acid composition. The most discriminating set of seven features were identified and successfully used to classify the effectors, with a sensitivity and specificity of over 80%. In addition, interdependence among the most discriminating seven features was revealed by graphical modeling. The most discriminating region for the most discriminating seven features was determined by sliding window analysis.

2. Materials and Methods

2.1 Dataset

We collected the 57 currently available amino acid sequences of symbiotic T3SS effector proteins from the literature (9, 15), and the same number of amino acid sequences for pathogenic T3SS effector proteins (9).

For each effector protein amino acid sequence, we calculated the physicochemical features, 41 in total, of charge, isoetectric point, number of proteolytic enzyme or reagent cleavage sites, mole percentage of each amino acid and amino acid groups defined in EMBOSS (16), and signal peptide probability. The list of 41 physicochemical features used in this study is in Table 1. Signal peptide probability was calculated by SignalP 3.0 (17), and others features were calculated by EMBOSS (16). These were used as attributes in our classification analysis.

2.2 Feature selection

We first used the Lepage test for the location-dispersion difference between the two groups (18). The top 10 discriminating features were chosen by the order of their p-values in the test statistics. The p-values of all of these candidate features were less than 0.001.

For these candidate features, we examined all combinations, 2^{10}-1, as explanatory variables in the kernel logistic regression (KLR), which is one of the kernel-learning methods suitable for binary-pattern recognition problems (19, 20). Let y_i be a binary observed

Table 1. Biochemical features used as attributes of effector proteins

No.	Description
1	Number of potentially antigenic regions of a protein sequence[1]
2	Number of proteolytic enzyme or reagent cleavage sites[1]
3	Number of secondary structure[1]
4	Hydrophobic moment[1]
5	Average residue weight[1]
6	Charge[1]
7	Isoelectric point[1]
8	Molar extinction coefficient[1]
9	Extinction coefficient at 1 mg/ml[1]
10	Probability of protein expression in E. coli inclusion bodies[1]
11-30	Mole percentage of each amino acid[1] 11:Ala, 12:Cys, 13:Asp, 14:Glu, 15:Phe, 16:Gly, 17:His, 18:Ile, 19:Lys, 20:Leu, 21:Met, 22:Asn, 23:Pro, 24:Gln, 25:Arg, 26:Ser, 27:Thr, 28:Val, 29:Trp, 30:Tyr
31	Mole percentage of tiny amino acids[1] (A+C+G+S+T)
32	Mole percentage of small amino acids[1] (A+B+C+D+G+N+P+S+T+V)
33	Mole percentage of aliphatic amino acids[1] (A+I+L+V)
34	Mole percentage of aromatic amino acids[1] (F+H+W+Y)
35	Mole percentage of non-polar amino acids[1] (A+C+F+G+I+L+M+P+V+W+Y)
36	Mole percentage of polar amino acids[1] (D+E+H+K+N+Q+R+S+T+Z)
37	Mole percentage of charged amino acids[1] (B+D+E+H+K+R+Z)
38	Mole percentage of basic amino acids[1] (H+K+R)
39	Mole percentage of acidic amino acids[1] (B+D+E+Z)
40	Number of cleavage sites between signal sequence and mature exported protein[1]
41	Signal peptide probability[2]

[1] calculated by EMBOSS (16). [2] calculated by SignalP (17).

variable and $p(\mathbf{x}_i)$ be its conditional distribution given $\mathbf{x}_i$, then the likelihood function was given by

$$L = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1-p(\mathbf{x}_i))^{1-y_i} \tag{1}$$

and log-likelihood function became

$$\begin{aligned}&\log L \\ &= \sum_{i=1}^{n} y_i \log \frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} + \log(1-p(\mathbf{x}_i))\end{aligned} \tag{2}$$

in which the unknown quantity $p(\mathbf{x}_i)$ was modeled using the radial basis kernel function $K(\mathbf{x}_j, \mathbf{x}_i)$ as

$$f(\mathbf{x}_i) = \log \frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} = \sum_{j=0}^{n} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) \tag{3}$$

where

$$\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \| \mathbf{x}_i - \mathbf{x}_j \|^2) \tag{4}$$

and σ is the kernel parameter. The solution of the parameter vector $\hat{\boldsymbol{\alpha}}$ was calculated using the following penalized log-likelihood function

$$\frac{1}{n}\{\sum_{i=1}^{n} y_i f(\mathbf{x}_i) - \log[1+\exp(f(\mathbf{x}_i))]\} - \frac{\lambda}{2}\boldsymbol{\alpha}^T \mathbf{R} \boldsymbol{\alpha} \tag{5}$$

where

$$R = \begin{pmatrix} 1 & \mathbf{1}_n^{\,T} \\ \mathbf{1}_n & \mathbf{K} \end{pmatrix}$$

by Fisher's scoring methods.

To select the best combination of the 10 candidate features , we used a generalized Bayesian information criterion (GBIC) (21). Using the likelihood function $L(\boldsymbol{\alpha})$ in equation (1) and the multivariate normal prior density $\pi(\boldsymbol{\alpha} \mid \lambda)$ for the parameter vector $\boldsymbol{\alpha}$ defined by

$$\pi(\boldsymbol{\alpha} \mid \lambda) = (2\pi)^{-r/2}(n\lambda)^{r/2} \mid R \mid^{1/2}_{+} \exp(-\frac{n\lambda}{2}\boldsymbol{\alpha}^T R \boldsymbol{\alpha}) \tag{6}$$

GBIC was defined as

$$GBIC = -2\log \int L(\boldsymbol{\alpha})\pi(\boldsymbol{\alpha} \mid \lambda) d\boldsymbol{\alpha} \tag{7}$$

and R was the same as that of equation (5), r was the rank of R, and $\mid R \mid_{+}$ was the product of r nonzero eigenvalues of R. Once $\hat{\boldsymbol{\alpha}}$ was obtained, GBIC was calculated through the Laplace approximation

$$\begin{aligned}&-2\log \int \exp(nl_\lambda(\boldsymbol{\alpha}))d\boldsymbol{\alpha} \\ &= -2\log\{\frac{(2\pi/n)^{(n+1)/2}}{\mid J_\lambda(\hat{\boldsymbol{\alpha}}) \mid^{1/2}} \exp(nl_\lambda(\hat{\boldsymbol{\alpha}}))\}\{1+O(n^{-1})\}\end{aligned} \tag{8}$$

where

$$l_\lambda(\boldsymbol{\alpha}) = \frac{1}{n}\log L(\boldsymbol{\alpha}) + \frac{1}{n}\log \pi(\boldsymbol{\alpha} \mid \lambda)$$

$$J_\lambda(\hat{\boldsymbol{\alpha}}) = -\frac{\partial^2 l_\lambda(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T},$$

GBIC was computed for each combination of 10 features , and the combination with the minimum GBIC was determined as explanatory variable of KLR. During the feature selection, values of kernel parameter σ and hyper parameter λ were given in the range of 1E-3 to 1E+3 (σ) or to 1E+4 (λ) for each set of explanatory features.

2.3 Classification performance

Classification performance using discriminating features identified by GBIC of KRL was analyzed using SVM based on the approximate relationship between KRL and the SVM (19). To determine the advantage of the most discriminating features, a misclassification rate was evaluated by leave-one-out cross-validation for each combination of k-features that attained the minimum GBIC in ${}_{10}C_k$ combinations (k=1,…,10). The results are summarized Figure 2, which illustrates the misclassification rates, with the number of features on the horizontal axis. We used svm function of e1071 package (E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel) in R.

2.4 Graphical modeling

To explore interdependence among the most discriminating features identified by GBIC of KLR, we used graphical modeling developed by Imoto et al. (22, 23) which combines non-linear nonparametric regression with radial basis and Bayesian network, and was originally developed for estimating genetic networks and functional relationships between genes. Non-linear nonparametric regression enabled us to capture directed dependencies among the features without advance knowledge about their relationships. Bayesian network is a powerful, graph-theoretic approach for expressing interdependence among variables as networks.

Calculations were conducted by MATLAB R2008b (The Mathworks Inc.) based on NETLAB (24), the Bayes net toolbox (BNT) for Matlab (25), and BNT structure learning package (26).

2.5 Sliding window analysis

N-terminal regions from the 1st to 97th residue were analyzed, with the window size varying from 8-50, and the starting position varying from 1 to 50. For each window, a dataset of the most discriminating features was created, and classification was conducted by SVM.

3. Results

3.1 Identification of discriminating features

A plot of minimum GBIC for ${}_{10}C_k$ combination of features used in KLR was given in Figure 1 taking the number of features, k, on the horizontal axis. The figure shows that the minimum GBIC tends to decrease as the number of features increase, take the smallest value when the number of features is seven, and increase at greater than seven features. The seven features that attained the smallest minimum GBIC were as follows: average residual weight, mole percentage of Ala, Asp, Ile, tiny amino acids, small amino acids, and acidic amino acids.

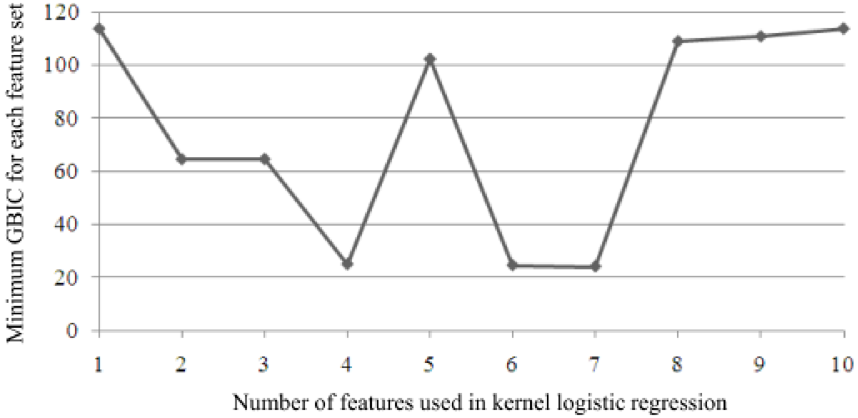

Figure 1. Plot of minimum GBIC against number of features used in kernel logistic regression.

3.2 Classification performance using the most discriminating features

Misclassification rates using the discriminating features identified by GBIC of KLR are plotted in Figure 2, taking the number of features on horizontal axis. The plot of minimum GBICs (Figure 1) and misclassification rates showed parallel tendencies. The best classification performance (84.2%) was obtained using a combination of the seven features that gave the smallest minimum GBIC (Figure 2A). The best performance with seven features was nearly identical to the results obtained when all 41 features were used. The seven discriminating features had a specificity of 85.5% and a sensitivity of 83.1% (Figure 2B).

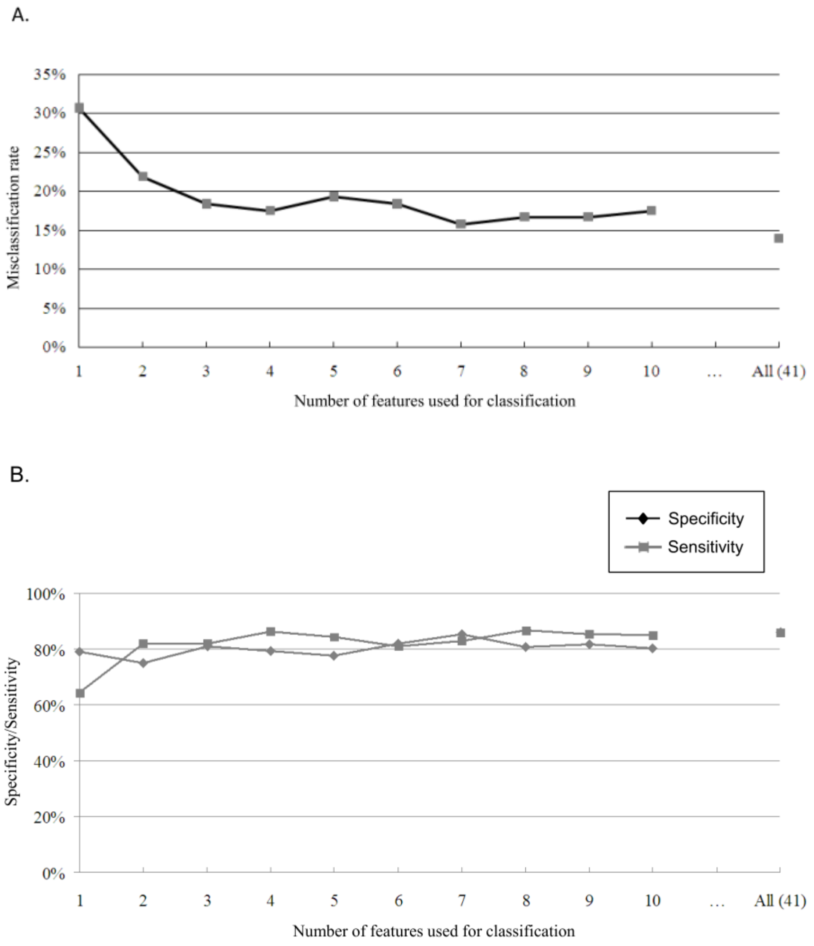

Figure 2. Classification performance using the discriminating features identified by GBIC of KLR. Misclassification rate for each combination of k-features that attained the minimum GBIC in $_{10}C_k$ combinations (k=1,…,10). Classification using all 41 features was also conducted, and the misclassification rate is at "All (41)" of the x-axis. **(A) Misclassification rate. (B) Specificity and sensitivity.**

3.3 Interdependence among the most discriminating features as a graph structure

The interdependence among the seven most discriminating features was represented in a directed-graph structure (Figure 3), in which the mole percentage of isoleucine, and a combination of alanine and average residue weight were positioned at the bottom end. The three features are representative of the directed-graph structure and have been selected by KLR at one or two features. Figure 2 shows that classification accuracy was about 70% for the mole percentage of isoleucine, and nearly 80% for a combination of alanine and average residue weight

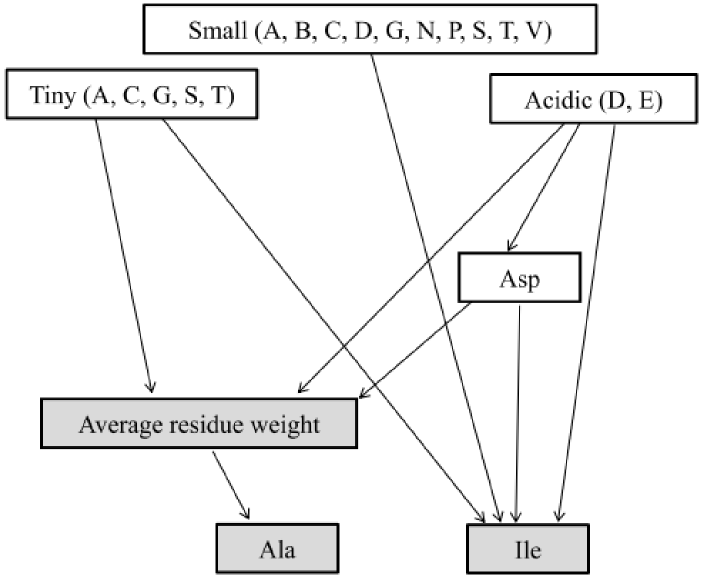

Figure 3. Graph structure showing interdependence among the most-discriminating features. Directed dependencies detected by nonparametric regression are depicted by arrows whose heads indicate response variables and tails indicate explanatory variables. Colours are the discriminating features identified by GBIC, when the number of features is one or two.

3.4 Identification of the most discriminating region

Sliding window analysis with variable window sizes and starting points is in Table 2. The region that gave the highest discrimination among the seven most-discriminating features was 48-95 residues from the N-terminus (N48-95), which gave a classification accuracy of 83.3% (Figure 4). Almost all of the second and third most-discriminating regions overlapped this region, supporting the hypothesis that the discriminating signature between pathogenic and symbiotic T3SS effector proteins was in this region.

Table 2. Results of sliding window analysis

Region	Misclassification rate	Starting point	Window size
N48-95	0.167	48	48
N49-95	0.175	49	47
N48-93	0.184	48	46
N48-96	0.184	48	49
N49-89	0.184	49	41
N49-90	0.184	49	42
N49-96	0.184	49	48
N9-36	0.184	9	28
N40-89	0.193	40	50
N47-93	0.193	47	47
N47-96	0.193	47	50
N48-92	0.193	48	45
N48-94	0.193	48	47
N49-93	0.193	49	45
N50-96	0.193	50	47
N65-97	0.193	65	33

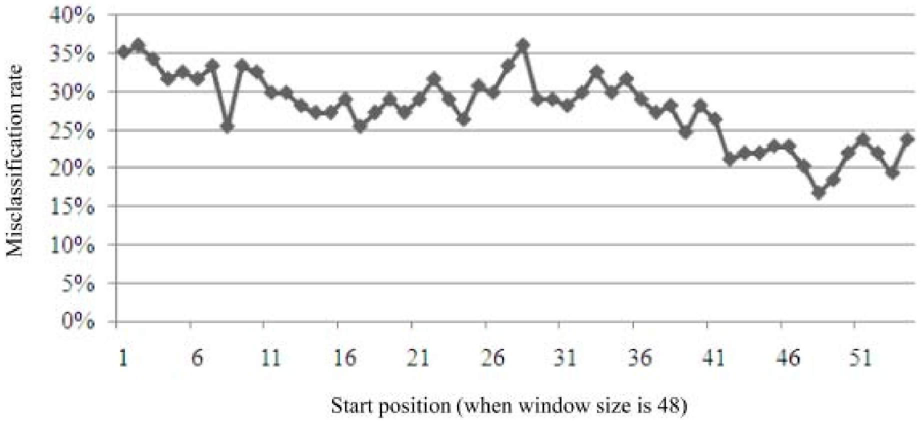

Figure 4. Plot of misclassification rate by sliding window analysis with window size 48. As shown in Table 2, misclassification rate is lowest when the analysis start position is 48 (*i.e.* for region N48 -95), and when the window size is 48, which gives the best classification performance.

3.5 Directions of differences of the discriminating features

The differences of the seven most discriminating features between pathogenic and symbiotic T3SS effector proteins are in Table 3, with "+ " meaning " more common in symbiotic proteins". Results are given for all regions, and for the most discriminating region, N48-95. The patterns of differences were almost equivalent between all regions and the most-discriminating region, supporting the hypothesis that N48-95 was the representative region that distinguished between pathogenic and symbiotic T3SS effector proteins. By mole percentage of amino acids, isoleucine decreased in symbiotic proteins, while the other amino acids (alanine, aspartic acid, acidic amino acids, tiny amino acid, small amino acid) increased in symbiotic proteins. The tendency was found both in all regions, and in the most discriminating N48-95 region.

Table 3. Directions of the differences of the most discriminating features

Pa	thogen		symbiont		direction*	
	(all regions)		(all regions)			
Feature M	ean	SD	mean	SD	Mean	SD
Ile (Molar %)	5.74	2.25	3.98	1.52	-	-
Average residue weight	109.30	4.12	108.98	2.80	-	-
Ala (Molar %)	8.28	3.07	10.99	2.80	+	-
Asp (Molar %)	4.49	1.91	16.01	1.60	+	-
Acidic (Molar %)	10.79	3.91	11.70	2.21	+	-
Tiny (Molar %)	31.58	6.98	32.94	3.90	+	-
Small (Molar %)	51.97	6.95	54.53	4.41	+	-
pathogen			symbiont		direction*	
	(N48-95)		(N48-95)			
Feature M	ean	SD	mean	SD	Mean	Region
Ile (Molar %)	5.30	4.09	3.84	2.81	-	-
Average residue weight	109.26	6.41	109.79	4.34	+	-
Ala (Molar %)	9.06	4.55	10.78	5.25	+	+
Asp (Molar %)	3.07	2.27	5.88	3.46	+	+
Acidic (Molar %)	8.92	5.69	11.15	4.91	+	-
Tiny (Molar %)	32.35	10.64	33.08	8.05	+	-
Small (Molar %)	51.68	10.95	53.91	6.69	+	-

* from pathogenic to symbiotic ("+" means "more in symbiotic proteins")

4. Discussion

In this work, we identified the seven most-discriminating features between pathogenic and symbiotic T3SS effector proteins, using a large combination of physicochemical features, analyzed by GBIC of KLR. The identified features were successfully used to classify the proteins by SVM, with sensitivities and specificities of over 80%.

The seven most-discriminating features were those related to amino acid composition. No other higher-order information was found to be as discriminating by GBIC of KLR. Interestingly, recently reported common features of T3SS effectors were also found to be amino acid composition or shared sequence motif. Embedded features in the amino acid sequence or composition may be a characteristic of T3SS effector proteins.

The most discriminating region between pathogenic and symbiotic effector proteins was 48-95 residues from the N-terminus. The classic signal peptide secretion signal is 15-40 residues from the N-terminus (27). Common features of T3SS effectors proteins were recently found to be embedded in 30 (10) or up to 50 residues (9) at the N-terminus. These findings are complementary with ours because the differences between pathogenic and symbiotic effector proteins are thought to have arisen after the common features in the N-terminus. Although common features are conserved, differences in amino acid composition occur, presumably because of different environments of pathogens, or symbiotic relationships with their hosts.

The identified discriminating features were used for classification, and for elucidating their interdependence using graphical modeling that combined non-linear nonparametric regression and Bayesian network. Although these techniques are usually used for estimating gene networks from microarray expression data, the combination of them, with feature selection, was a powerful method for a deeper understanding of the meaning of the discriminating features.

This is the first study to explore discriminating features between pathogenic and symbiotic T3SS effector proteins, using a combination of computational statistical and machine-learning approaches. The most-discriminating features, their interdependence, and the most-discriminating region were determined by these methods. This study will provide a methodological basis for future research, and provides important insight about the functional differences between pathogenic and symbiotic T3SS effectors.

Acknowledgement The computational calculations were carried out at the Human Genome Center at the Institute of Medical Science, the University of Tokyo. This work was supported by a grant from the Science and Technology Foundation of Japan to Koji Yahara.

Referecnes

1. Cornelis, G. R. (2006) *Nat Rev Microbiol* **4,** 811-25.
2. Coburn, B., Sekirov, I. & Finlay, B. B. (2007) *Clin Microbiol Rev* **20,** 535-49.
3. Hernandez, L. D., Pypaert, M., Flavell, R. A. & Galan, J. E. (2003) *J Cell Biol* **163,** 1123-31.
4. Boyle, E. C., Brown, N. F. & Finlay, B. B. (2006) *Cell Microbiol* **8,** 1946-57.
5. Yoshida, S., Katayama, E., Kuwae, A., Mimuro, H., Suzuki, T. & Sasakawa, C. (2002) *Embo J* **21,** 2923-35.
6. Beeckman, D. S. & Vanrompay, D. C. (2009) *Curr Issues Mol Biol* **12,** 17-42.
7. Coombes, B. K. (2009) *Trends Microbiol* **17,** 89-94.
8. Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G. J., Droege, M., Frishman, D., Rattei, T., Mewes, H. W. & Wagner, M. (2004) *Science* **304,** 728-30.
9. Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H. W., Horn, M. & Rattei, T. (2009) *PLoS Pathog* **5,** e1000376.
10. Samudrala, R., Heffron, F. & McDermott, J. E. (2009) *PLoS Pathog* **5,** e1000375.
11. Grynberg, M. & Godzik, A. (2009) *PLoS Pathog* **5,** e1000398.
12. Kambara, K., Ardissone, S., Kobayashi, H., Saad, M. M., Schumpp, O., Broughton, W. J. & Deakin, W. J. (2009) *Mol Microbiol* **71,** 92-106.
13. Masson-Boivin, C., Giraud, E., Perret, X. & Batut, J. (2009) *Trends Microbiol* **17,** 458-66.
14. Dale, C. & Moran, N. A. (2006) *Cell* **126,** 453-465.
15. Lower, M. & Schneider, G. (2009) *PLoS One* **4,** e5917.
16. Rice, P., Longden, I. & Bleasby, A. (2000) *Trends Genet* **16,** 276-7.
17. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004) *J Mol Biol* **340,** 783-95.
18. Lepage, Y. (1971) *Biometrika* **58,** 213-217.
19. Zhu, J. & Hastie, T. (2001) *Journal of Computational and Graphical Statistics* **14,** 1081-1088.
20. Cawley, G. C. & Talbot, N. L. (2008) *Machine Learning* **71,** 243 - 264.
21. Konishi, S., Ando, T. & Imoto, S. (2004) *Biometrika* **91,** 27-43.
22. Imoto, S., Goto, T. & Miyano, S. (2002) *Pac Symp Biocomput***,** 175-86.
23. Imoto, S., Sunyong, K., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. & Miyano, S. (2002) *Proc IEEE Comput Soc Bioinform Conf* **1,** 219-27.
24. Nabney, I. (2001) *NETLAB: algorithms for pattern recognition* (Springer, London).
25. Murphy, K. P. (2001) *Computing Science and Statistics.* **33,** 331-350.
26. Leray, P. & Francois, O. (2006) (Laboratoire PSI, Universitè et INSA de Rouen.
27. von Heijne, G. (1985) *J Mol Biol* **184,** 99-105.